(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113011155 A
(43)申请公布日 2021.06.22

(21)申请号 202110282699.5

(22)申请日 2021.03.16

(71)申请人 北京百度网讯科技有限公司
地址 100094 北京市海淀区上地十街10号百度大厦2层

(72)发明人 马超 张敬帅 黄启帆 姚开春 王鹏 祝恒书

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 赵林琳

(51)Int.Cl.
G06F 40/194(2020.01)
G06F 40/289(2020.01)
G06F 40/30(2020.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
G06Q 10/10(2012.01)

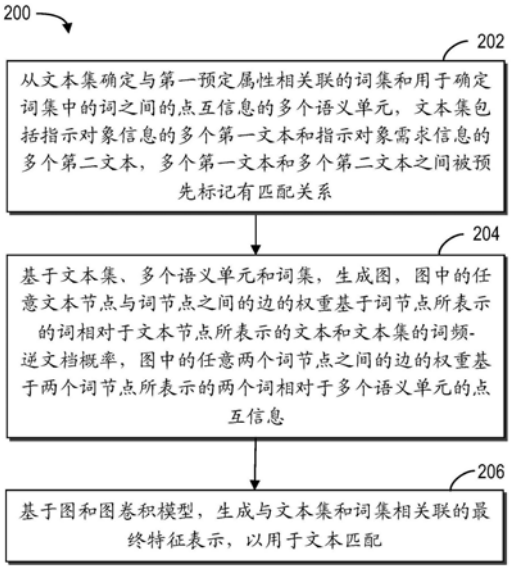
权利要求书3页 说明书9页 附图7页

(54)发明名称

用于文本匹配的方法、装置、设备、存储介质和程序产品

(57)摘要

本公开公开了用于文本匹配的方法、装置、设备、存储介质和程序产品，涉及计算机技术领域，尤其涉及自然语言处理和深度学习领域。具体实现方案为：从文本集确定与第一预定属性相关联的词集和多个语义单元，文本集包括指示对象信息的多个第一文本和指示对象需求信息的多个第二文本；生成图，图中的任意文本节点与词节点之间的边的权重基于词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档频率，图中的任意两个词节点之间的边的权重基于两个词节点所表示的两个词相对于多个语义单元的点互信息；以及基于图和图卷积模型，生成文本集和词集的最终特征表示，以用于文本匹配。由此，能够使得文本特征有更准确的语义表达能力。



CN 113011155 A

1.一种用于文本匹配的方法,包括:

从文本集确定与第一预定属性相关联的词集和用于确定所述词集中的词之间的点互信息的多个语义单元,所述文本集包括指示对象信息的多个第一文本和指示对象需求信息的多个第二文本,所述多个第一文本和所述多个第二文本之间被预先标记有匹配关系;

基于所述文本集、所述多个语义单元和所述词集,生成图,所述图中的任意文本节点与词节点之间的边的权重基于所述词节点所表示的词相对于所述文本节点所表示的文本和所述文本集的词频-逆文档频率,所述图中的任意两个词节点之间的边的权重基于所述两个词节点所表示的两个词相对于所述多个语义单元的点互信息;以及

基于所述图和图卷积模型,生成与所述文本集和所述词集相关联的最终特征表示,以用于文本匹配。

2.根据权利要求1所述的方法,其中确定所述多个语义单元包括以下至少一项:

确定所述第一文本中与第二预定属性相关联的文本片段和被预先标记为与所述第一文本相匹配的第二文本中的对象需求信息片段,作为同一个语义单元;

确定所述第一文本中与所述第二预定属性相关联的文本片段作为语义单元;以及

确定所述第二文本中的对象需求信息片段作为语义单元。

3.根据权利要求1所述的方法,其中生成所述图包括:

如果确定词节点所表示的词在文本节点所表示的文本中出现,则:

生成所述词节点与所述文本节点之间的边;以及

确定所述词节点所表示的词相对于所述文本节点所表示的文本和所述文本集的词频-逆文档频率,作为所述词节点与所述文本节点之间的边的权重。

4.根据权利要求1所述的方法,其中生成所述图包括:

在所述多个语义单元中确定包括任意两个词节点所表示的两个词中的第一词的语义单元的第一数量、包括所述两个词中的第二词的语义单元的第二数量以及同时包括所述两个词的语义单元的第三数量;

基于所述多个语义单元的总数量、所述第一数量、所述第二数量和所述第三数量,确定所述两个词之间的点互信息;以及

基于所述点互信息,生成所述两个词节点之间的边的权重。

5.根据权利要求4所述的方法,其中生成所述两个词节点之间的边的权重包括:

如果确定所述点互信息大于零,则:

生成所述两个词节点之间的边;以及

将所述点互信息确定为所述两个词节点之间的边的权重。

6.根据权利要求1所述的方法,其中生成所述最终特征表示包括:

基于所述图中的边的权重,生成邻接矩阵;

生成与所述文本集和所述词集相关联的初始特征表示;以及

基于所述邻接矩阵、所述初始特征表示和图卷积模型,生成与所述文本集和所述词集相关联的所述最终特征表示,以用于文本匹配。

7.根据权利要求6所述的方法,其中生成所述最终特征表示包括:

S1:基于所述邻接矩阵、所述初始特征表示和所述图卷积模型,生成与所述文本集和所述词集相关联的中间特征表示;

S2:基于所述中间特征表示和目标函数,通过梯度下降更新所述图卷积模型,所述目标函数用于最大化所述文本集中被预先标记为相匹配的第一文本和第二文本之间的特征相似度并且最小化与所述文本集中被预先标记为不匹配的第一文本和第二文本之间的特征相似度;以及

重复步骤S1和S2,直至所述目标函数的梯度收敛,以及将收敛时的中间特征表示作为所述最终特征表示。

8.根据权利要求1所述的方法,其中第一文本为简历文本,以及第二文本为岗位描述文本。

9.一种用于文本匹配的装置,包括:

词集与语义单元确定模块,用于从文本集确定与第一预定属性相关联的词集和用于确定所述词集中的词之间的点互信息的多个语义单元,所述文本集包括指示对象信息的多个第一文本和指示对象需求信息的多个第二文本,所述多个第一文本和所述多个第二文本之间被预先标记有匹配关系;

图生成模块,用于基于所述文本集、所述多个语义单元和所述词集,生成图,所述图中的任意文本节点与词节点之间的边的权重基于所述词节点所表示的词相对于所述文本节点所表示的文本和所述文本集的词频-逆文档频率,所述图中的任意两个词节点之间的边的权重基于所述两个词节点所表示的两个词相对于所述多个语义单元的点互信息;以及

特征表示生成模块,用于基于所述图和图卷积模型,生成与所述文本集和所述词集相关联的最终特征表示,以用于文本匹配。

10.根据权利要求9所述的装置,其中词集与语义单元确定模块还用于以下至少一项:

确定所述第一文本中与第二预定属性相关联的文本片段和被预先标记为与所述第一文本相匹配的第二文本中的对象需求信息片段,作为同一个语义单元;

确定所述第一文本中与所述第二预定属性相关联的文本片段作为语义单元;以及

确定所述第二文本中的对象需求信息片段作为语义单元。

11.根据权利要求9所述的装置,其中所述图生成模块还用于:

如果确定词节点所表示的词在文本节点所表示的文本中出现,则:

生成所述词节点与所述文本节点之间的边;以及

确定所述词节点所表示的词相对于所述文本节点所表示的文本和所述文本集的词频-逆文档频率,作为所述词节点与所述文本节点之间的边的权重。

12.根据权利要求9所述的装置,其中所述图生成模块包括:

语义单元数量确定子模块,用于在所述多个语义单元中确定包括任意两个词节点所表示的两个词中的第一词的语义单元的第一数量、包括所述两个词中的第二词的语义单元的第二数量以及同时包括所述两个词的语义单元的第三数量;

点互信息确定子模块,用于基于所述多个语义单元的总数量、所述第一数量、所述第二数量和所述第三数量,确定所述两个词之间的点互信息;以及

权重生成子模块,用于基于所述点互信息,生成所述两个词节点之间的边的权重。

13.根据权利要求12所述的装置,其中所述权重生成子模块还用于:

如果确定所述点互信息大于零,则:

生成所述两个词节点之间的边;以及

将所述点互信息确定为所述两个词节点之间的边的权重。

14.根据权利要求9所述的装置,其中所述特征表示生成模块包括:

邻接矩阵生成子模块,用于基于所述图中的边的权重,生成邻接矩阵;

初始特征表示生成子模块,用于生成与所述文本集和所述词集相关联的初始特征表示;以及

最终特征表示生成子模块,用于基于所述邻接矩阵、所述初始特征表示和图卷积模型,生成与所述文本集和所述词集相关联的所述最终特征表示,以用于文本匹配。

15.根据权利要求14所述的装置,其中所述最终特征表示生成子模块还用于:

S1:基于所述邻接矩阵、所述初始特征表示和所述图卷积模型,生成与所述文本集和所述词集相关联的中间特征表示;

S2:基于所述中间特征表示和目标函数,通过梯度下降更新所述图卷积模型,所述目标函数用于最大化所述文本集中被预先标记为相匹配的第一文本和第二文本之间的特征相似度并且最小化与所述文本集中被预先标记为不匹配的第一文本和第二文本之间的特征相似度;以及

重复步骤S1和S2,直至所述目标函数的梯度收敛,以及将收敛时的中间特征表示作为所述最终特征表示。

16.根据权利要求9所述的装置,其中第一文本为简历文本,以及第二文本为岗位描述文本。

17.一种电子设备,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-8中任一项所述的方法。

18.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使所述计算机执行根据权利要求1-8中任一项所述的方法。

19.一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时实现根据权利要求1-8中任一项所述的方法。

用于文本匹配的方法、装置、设备、存储介质和程序产品

技术领域

[0001] 本公开涉及计算机技术领域，尤其涉及自然语言处理和深度学习技术。

背景技术

[0002] 在文本匹配的应用领域中，诸如简历与岗位匹配中，通常利用人工来分析简历信息和岗位描述信息，并根据个人经验对简历和岗位的匹配程度进行判断，效率低下并且带有个人主观色彩。

发明内容

[0003] 本公开提供了一种用于文本匹配的方法、装置、设备、存储介质和程序产品。

[0004] 根据本公开的第一方面，提供了一种用于文本匹配的方法，包括：从文本集确定与第一预定属性相关联的词集和用于确定词集中的词之间的点互信息的多个语义单元，文本集包括指示对象信息的多个第一文本和指示对象需求信息的多个第二文本，多个第一文本和多个第二文本之间被预先标记有匹配关系；基于文本集、多个语义单元和词集，生成图，图中的任意文本节点与词节点之间的边的权重基于词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档频率，图中的任意两个词节点之间的边的权重基于两个词节点所表示的两个词相对于多个语义单元的点互信息；以及基于图和图卷积模型，生成与文本集和词集相关联的最终特征表示，以用于文本匹配。

[0005] 根据本公开的第二方面，提供了一种用于文本匹配的装置，包括：词集与语义单元确定模块，用于从文本集确定与第一预定属性相关联的词集和用于确定词集中的词之间的点互信息的多个语义单元，文本集包括指示对象信息的多个第一文本和指示对象需求信息的多个第二文本，多个第一文本和多个第二文本之间被预先标记有匹配关系；图生成模块，用于基于文本集、多个语义单元和词集，生成图，图中的任意文本节点与词节点之间的边的权重基于词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档频率，图中的任意两个词节点之间的边的权重基于两个词节点所表示的两个词相对于多个语义单元的点互信息；以及特征表示生成模块，用于基于图和图卷积模型，生成与文本集和词集相关联的最终特征表示，以用于文本匹配。

[0006] 根据本公开的第三方面，提供了一种电子设备，包括：至少一个处理器；以及与所述至少一个处理器通信连接的存储器；其中，所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行根据第一方面所述的方法。

[0007] 根据本公开的第四方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，其中，所述计算机指令用于使所述计算机执行根据第一方面所述的方法。

[0008] 根据本公开的第五方面，提供了一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现根据第一方面所述的方法。

[0009] 由此，能够捕捉全局词语的共现信息以及文本和单词之间的关系，使得文本特征

表示有更准确的语义表达能力。

[0010] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0011] 附图用于更好地理解本方案，不构成对本公开的限定。

[0012] 图1是根据本公开实施例的信息处理环境100的示意图。

[0013] 图2是根据本公开实施例的用于文本匹配的方法200的示意图。

[0014] 图3是根据本公开的实施例的用于生成图的方法300的流程图。

[0015] 图4是根据本公开的实施例的用于生成图的方法400的流程图。

[0016] 图5是根据本公开的实施例的用于生成最终特征表示的方法500的流程图。

[0017] 图6是根据本公开的实施例的用于文本匹配的装置600的示意图。

[0018] 图7是根据本公开的实施例的图700的示意图。

[0019] 图8是用来实现本公开实施例的用于文本匹配的方法的电子设备的框图。

具体实施方式

[0020] 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0021] 如上所述，在利用人工文档匹配的应用领域中，例如利用人工匹配人员简历和工作岗位，效率低下且带有个人主观色彩。在一种传统方案中，可以通过textCNN(文本卷积神经网络)对简历和岗位描述文本进行表征，但是textCNN难以捕捉全局词语的共现信息和文本与单词之间的关系，使得特征表示的语义表达能力不够准确。

[0022] 为了至少部分地解决上述问题以及其他潜在问题中的一个或者多个，本公开的示例实施例提出了一种用于文本匹配的方案。在该方案中，计算设备从文本集确定与第一预定属性相关联的词集和用于确定词集中的词之间的点互信息的多个语义单元，文本集包括指示对象信息的多个第一文本和指示对象需求信息的多个第二文本，多个第一文本和多个第二文本之间被预先标记有匹配关系。随后，计算设备基于文本集、多个语义单元和词集，生成图，图中的任意文本节点与词节点之间的边的权重基于词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档频率，图中的任意两个词节点之间的边的权重基于两个词节点所表示的两个词相对于多个语义单元的点互信息。接着，计算设备基于图和图卷积模型，生成与文本集和词集相关联的最终特征表示，以用于文本匹配。以此方式，能够捕捉全局词语的共现信息以及文本和单词之间的关系，使得文本特征表示有更准确的语义表达能力。

[0023] 在下文中，将结合附图更详细地描述本公开的具体实施例。

[0024] 图1示出了根据本公开的实施例的信息处理环境100的示例的示意图。信息处理环境100可以包括计算设备110、文本集120和最终特征表示130。

[0025] 计算设备110例如包括但不限于服务器计算机、多处理器系统、大型计算机、包括

上述系统或设备中的任意一个的分布式计算环境等。在一些实施例中，计算设备110可以具有一个或多个处理单元，包括诸如图像处理单元GPU、现场可编程门阵列FPGA和专用集成电路ASIC等的专用处理单元以及诸如中央处理单元CPU的通用处理单元。

[0026] 文本集120包括指示对象信息的多个第一文本121-1至121-n(以下对“第一文本”统称为121)和指示对象需求信息的多个第二文本122-1至122-n(以下对“第二文本”统称为122)。多个第一文本121和多个第二文本122之间被预先标记有匹配关系。例如，一个第一文本121可以被预先标记为与一个或多个第二文本122相匹配，或者不匹配。文本集120中的第一文本和第二文本可以包括与第一预定属性相关联的词。例如在个人简历与岗位描述的场景中，与第一预定属性相关联的词可以包括但不限于与技能相关联的词。

[0027] 计算设备110用于从文本集120确定与第一预定属性相关联的词集和用于确定词集中的词之间的点互信息的多个语义单元。文本集120包括指示对象信息的多个第一文本121和指示对象需求信息的多个第二文本122，多个第一文本121和多个第二文本122之间被预先标记有匹配关系。计算设备110基于文本集120、多个语义单元和词集，生成图。该图中的任意文本节点与词节点之间的边的权重基于词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档频率。该图中的任意两个词节点之间的边的权重基于两个词节点所表示的两个词相对于多个语义单元的点互信息。并且计算设备110基于图和图卷积模型，生成与文本集120和词集相关联的最终特征表示130，以用于文本匹配。

[0028] 由此，能够捕捉全局词语的共现信息以及文本和单词之间的关系，使得文本特征表示有更准确的语义表达能力。

[0029] 图2示出了根据本公开的实施例的用于文本匹配的方法200的流程图。例如，方法200可以由如图1所示的计算设备110来执行。应当理解的是，方法200还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。

[0030] 在框202处，计算设备110从文本集120确定与第一预定属性相关联的词集和用于确定词集中的词之间的点互信息的多个语义单元。文本集120包括指示对象信息的多个第一文本121和指示对象需求信息的多个第二文本122，多个第一文本121和多个第二文本122之间被预先标记有匹配关系。

[0031] 例如，在个人简历和岗位描述的应用领域中，第一文本121例如可以是简历文本，以及第二文本122可以是岗位描述文本。第一预定属性例如包括但不限于技能。计算设备110可以从多个简历文本和多个简历描述文本确定与技能相关联的词集。应当理解，虽然文中以个人简历和岗位描述的应用领域进行举例说明，但本公开的范围在此不受限制，本公开也可以应用于其他文本匹配领域。

[0032] 在一些实施例中，计算设备110可以确定第一文本121中与第二预定属性相关联的文本片段作为语义单元。例如，在个人简历和岗位描述的应用领域中，第二预定属性例如包括但不限于工作经历。计算设备110可以确定简历文本中用于描述一段工作经历的文本片段作为一个语义单元。

[0033] 备选地或者附加地，在一些实施例中，计算设备110可以确定第二文本122中的对象需求信息片段作为语义单元。例如，在个人简历和岗位描述的应用领域中，计算设备110可以确定岗位描述文本中的一句或者一段需求描述作为一个语义单元。

[0034] 备选地或者附加地，在一些实施例中，计算设备110可以确定第一文本121中与第

二预定属性相关联的文本片段和被预先标记为与该第一文本121相匹配的第二文本122中的对象需求信息片段，作为同一个语义单元。例如，在个人简历和岗位描述的应用领域中，某个简历文本被预先标记为与某个岗位描述文本相匹配，则计算设备110可以将该简历文本中用于描述一段工作经历的文本片段和该岗位描述文本中的需求描述部分作为同一个语义单元。

[0035] 由此，能够确定用于对词共现进行统计的语义单元，便于捕获全局词共现信息。

[0036] 在框204处，计算设备110基于文本集120、多个语义单元和词集，生成图，图中的任意文本节点与词节点之间的边的权重基于词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档频率，图中的任意两个词节点之间的边的权重基于两个词节点所表示的两个词相对于多个语义单元的点互信息。

[0037] 所生成的图的示例可如图7所示。其中，圆形节点710-750表示词节点，方形节点760-790表示文本节点。应当理解，图7只是举例说明，所生成的图中的节点和节点之间的边不限于图7所示的情形。下文将结合图3和图4详细描述用于生成图的方法。

[0038] 继续对图2的描述，在框206处，计算设备110基于图和图卷积模型，生成与文本集和词集相关联的最终特征表示130，以用于文本匹配。下文将结合图5详细描述用于生成最终特征表示130的方法。

[0039] 由此，根据本公开的实施方式能够捕捉全局词语的共现信息以及文本和单词之间的关系，使得文本特征表示有更准确的语义表达能力。此外，为了剔除其余信息的干扰，只考虑简历文本和岗位描述文本中的技能信息对其进行表征，通过图卷积模型可以准确捕捉技能词之间的层次关系。还有，图卷积模型对于不同场景或需求，可以灵活定义优化目标，得到更符合场景或需求的特征表示。

[0040] 图3示出了根据本公开的实施例的用于生成图的方法300的流程图。例如，方法300可以由如图1所示的计算设备110来执行。应当理解的是，方法300还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。

[0041] 在框302处，计算设备110确定词节点所表示的词在文本节点所表示的文本中是否出现。

[0042] 如果在框302处计算设备110确定词节点所表示的词在文本节点所表示的文本中出现，则在框304处生成词节点与文本节点之间的边。

[0043] 在框306处，计算设备110确定词节点所表示的词相对于文本节点所表示的文本和文本集120的词频-逆文档频率，作为词节点与文本节点之间的边的权重。

[0044] 具体来说，计算设备110可以确定词节点所表示的词在文本节点所表示的文本中的出现次数作为词频，以及确定文本集120中包括该词的文本数，并基于文本数和文本集120中的总文本数，确定逆文档频率，从而确定词频-逆文档频率(TF-IDF)。

[0045] 由此，能够捕获文本集中的文本与词集中的词之间的关系，便于生成具有更准确语义表达的特征表示。

[0046] 图4示出了根据本公开的实施例的用于生成图的方法400的流程图。例如，方法400可以由如图1所示的计算设备110来执行。应当理解的是，方法400还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。

[0047] 在框402处，计算设备110在多个语义单元中确定包括任意两个词节点所表示的两

个词中的第一词的语义单元的第一数量、包括两个词中的第二词的语义单元的第二数量以及同时包括两个词的语义单元的第三数量。

[0048] 在框404处，计算设备110基于多个语义单元的总数量、第一数量、第二数量和第三数量，确定两个词之间的点互信息。

[0049] 例如，可以通过以下公式1-4来确定两个词i和j之间的点互信息。

[0050] $$\mathrm{PMI}(i,j)=log\frac{p(i,j)}{p(i)p(j)} \quad 1)$$

[0051] $$p(i,j)=\frac{\#W(i,j)}{\#W} \quad 2)$$

[0052] $$p(i)=\frac{\#W(i)}{\#W} \quad 3)$$

[0053] $$p(j)=\frac{\#W(j)}{\#W} \quad 4)$$

[0054] 其中，#W表示多个语义单元的总数量，#W(i)表示包括词i的语义单元的第一数量，#W(j)表示包括词j的语义单元的第二数量，#W(i,j)表示同时包括词i和词j的语义单元的第三数量。

[0055] 在框406处，计算设备110基于点互信息，生成两个词节点之间的边的权重。例如，计算设备110可以将两个词之间的点互信息作为两个词节点之间的边的权重。

[0056] 在一些实施例中，计算设备110可以确定两个词之间的点互信息是否大于零。如果计算设备110确定点互信息大于零，则生成两个词节点之间的边，以及将两个词之间的点互信息确定为两个词节点之间的边的权重。由于点互信息为正表示词与词之间的语义相关性较高，为负表示两个词之间的语义相关性较低或不存在，通过给点互信息为正的两个词节点生成边，有利于捕捉语义相关性较高的词。

[0057] 由此，能够捕获两个词相对于文本集中多个语义单元的全局共现信息，便于生成具有更准确语义表达的特征表示。

[0058] 图5示出了根据本公开的实施例的用于生成最终特征表示的方法500的流程图。例如，方法500可以由如图1所示的计算设备110来执行。应当理解的是，方法500还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。

[0059] 在框502处，计算设备110基于图中的边的权重，生成邻接矩阵。

[0060] 例如，邻接矩阵A中的元素A_{ij}可以通过以下公式5来表示。

[0061] $$A_{ij}=\begin{cases}PMI(i,j) & i,j\text{都是词},PMI(i,j)>0\\ TF\,IDF_{ij} & i\text{是文本，j 是词}\\ 1 & i=j\\ 0 & \text{其他}\end{cases} \quad 5)$$

[0062] 在框504处，计算设备110生成与文本集和词集相关联的初始特征表示。

[0063] 例如，计算设备110可以分别生成多个第一文本的多个第一词表示、多个第二文本的多个第二词表示以及词集的多个第三词表示，并且将多个第一词表示、多个第二词表示

和多个第三词表示拼接，生成初始特征表示。

[0064] 在框506处，计算设备110基于邻接矩阵、初始特征表示和图卷积模型，生成与文本集120和词集相关联的最终特征表示，以用于文本匹配。

[0065] 例如，例如可以采用二层图卷积模型，可通过以下公式6来表示。

[0066] $$Z = Softmax(\tilde{A}ReLU(\tilde{A}XW_0)W_1) \quad (6)$$

[0067] 其中，X为初始特征表示，W_0和W_1为模型参数，Z为最终特征表示，$\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$，D为邻接矩阵A的度矩阵，$D_{ii} = \sum_j A_{ij}$。第一层可以使用ReLU作为激活函数，第二层对输出使用softmax进行归一化输出。

[0068] 由此，能够通过图卷积模型来捕捉与第一预定属性相关联的词集之间的层次关系，便于生成具有更准确语义表达的文本特征表示。

[0069] 在一些实施例中，可以对图卷积模型的目标函数进行优化，使得目标函数用于最大化文本集120中被预先标记为相匹配的第一文本和第二文本之间的特征相似度并且最小化与文本集120中被预先标记为不匹配的第一文本和第二文本之间的特征相似度。特征相似度例如包括但不限于余弦相似度，或者点乘。

[0070] 具体来说，计算设备110可以基于邻接矩阵、初始特征表示和图卷积模型，生成与文本集和所述词集相关联的中间特征表示。

[0071] 随后，计算设备110可以基于中间特征表示和目标函数，通过梯度下降更新图卷积模型，目标函数用于最大化文本集中被预先标记为相匹配的第一文本和第二文本之间的特征相似度并且最小化与文本集中被预先标记为不匹配的第一文本和第二文本之间的特征相似度。

[0072] 重复上述两个步骤，直至目标函数的梯度收敛，以及将收敛时的中间特征表示作为最终特征表示。

[0073] 作为示例，目标函数可以通过以下公式7来实现。

[0074] $$\min(-\sum_{i=1}^{|M|}\cos(j_i, r_i) + \sum_{f=1}^{|F|}\cos(j_f, r_f)) \quad (7)$$

[0075] 其中，M表示文本集120中被预先标记为相匹配的第一文本和第二文本的集合，i表示该集合中的第i对相匹配的第一文本和第二文本，F表示文本集120中被预先表示为不匹配的第一文本和第二文本的集合，f表示该集合中的第f对不匹配的第一文本和第二文本，r表示第一文本，j表示第二文本。例如，在个人简历和岗位描述的应用领域中，M表示文本集120中被预先标记为相匹配的简历文本和岗位描述文本的集合，i表示该集合中的第i对相匹配的简历文本和岗位描述文本，F表示文本集120中被预先表示为不匹配的简历文本和岗位描述文本的集合，f表示该集合中的第f对不匹配的简历文本和岗位描述文本，r表示简历文本，j表示岗位描述文本。

[0076] 由此，根据本公开的实施方式能够使得通过图卷积模型生成的最终特征表示满足相匹配的第一文本和第二文本对应的余弦相似度最大化，不匹配的第一文本和第二文本对应的余弦相似度最小化，从而能够解决传统自动文档匹配方案中无法使用负样本进行建模的问题，以及无法使用自定义的优化目标缺乏灵活性的问题。

[0077] 图6示出了根据本公开的实施例的用于文本匹配的装置600的示意图。如图6所示，装置600包括词集与语义单元确定模块610、图生成模块620和特征表示生成模块630。

[0078] 关于词集与语义单元确定模块610，其用于从文本集120确定与第一预定属性相关联的词集和用于确定词集中的词之间的点互信息的多个语义单元，文本集120包括指示对象信息的多个第一文本和指示对象需求信息的多个第二文本，多个第一文本和多个第二文本之间被预先标记有匹配关系。

[0079] 关于图生成模块620，其用于基于文本集120、多个语义单元和词集，生成图，图中的任意文本节点与词节点之间的边的权重基于词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档频率，图中的任意两个词节点之间的边的权重基于两个词节点所表示的两个词相对于多个语义单元的点互信息。

[0080] 关于特征表示生成模块630，其用于基于图和图卷积模型，生成与文本集120和词集相关联的最终特征表示，以用于文本匹配。

[0081] 由此，能够捕捉全局词语的共现信息以及文本和单词之间的关系，使得文本特征表示有更准确的语义表达能力。

[0082] 在一些实施例中，词集与语义单元确定模块610还用于以下至少一项：确定第一文本中与第二预定属性相关联的文本片段和被预先标记为与第一文本相匹配的第二文本中的对象需求信息片段，作为同一个语义单元；确定第一文本中与第二预定属性相关联的文本片段作为语义单元；以及确定第二文本中的对象需求信息片段作为语义单元。

[0083] 备选地或者附加地，在一些实施例中，图生成模块620还用于如果确定词节点所表示的词在文本节点所表示的文本中出现，则生成词节点与文本节点之间的边；以及确定词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档频率，作为词节点与文本节点之间的边的权重。

[0084] 备选地或者附加地，在一些实施例中，图生成模块620包括语义单元数量确定子模块，用于在多个语义单元中确定包括任意两个词节点所表示的两个词中的第一词的语义单元的第一数量、包括两个词中的第二词的语义单元的第二数量以及同时包括两个词的语义单元的第三数量；点互信息确定子模块，用于基于多个语义单元的总数量、第一数量、第二数量和第三数量，确定两个词之间的点互信息；以及权重生成子模块，用于基于点互信息，生成两个词节点之间的边的权重。

[0085] 备选地或者附加地，在一些实施例中，权重生成子模块还用于如果确定点互信息大于零，则生成两个词节点之间的边；以及将点互信息确定为两个词节点之间的边的权重。

[0086] 备选地或者附加地，在一些实施例中，特征表示生成模块630包括邻接矩阵生成子模块，用于基于图中的边的权重，生成邻接矩阵；初始特征表示生成子模块，用于生成与文本集和词集相关联的初始特征表示；以及最终特征表示生成子模块，用于基于邻接矩阵、初始特征表示和图卷积模型，生成与文本集和词集相关联的最终特征表示，以用于文本匹配。

[0087] 备选地或者附加地，在一些实施例中，最终特征表示生成子模块还用于：S1：基于邻接矩阵、初始特征表示和图卷积模型，生成与文本集和词集相关联的中间特征表示；S2：基于中间特征表示和目标函数，通过梯度下降更新图卷积模型，目标函数用于最大化所述文本集中被预先标记为相匹配的第一文本和第二文本之间的特征相似度并且最小化与文本集中被预先标记为不匹配的第一文本和第二文本之间的特征相似度；以及重复步骤S1和

S2,直至目标函数的梯度收敛,以及将收敛时的中间特征表示作为最终特征表示。

[0088] 在一些实施例中,第一文本为简历文本,以及第二文本为岗位描述文本。

[0089] 根据本公开的实施例,本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0090] 图8示出了可以用来实施本公开的实施例的示例电子设备800的示意性框图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0091] 如图8所示,设备800包括计算单元801,其可以根据存储在只读存储器(ROM)802中的计算机程序或者从存储单元808加载到随机访问存储器(RAM)803中的计算机程序,来执行各种适当的动作和处理。在RAM 803中,还可存储设备800操作所需的各种程序和数据。计算单元801、ROM 802以及RAM 803通过总线804彼此相连。输入/输出(I/O)接口805也连接至总线804。

[0092] 设备800中的多个部件连接至I/O接口805,包括:输入单元806,例如键盘、鼠标等;输出单元807,例如各种类型的显示器、扬声器等;存储单元808,例如磁盘、光盘等;以及通信单元809,例如网卡、调制解调器、无线通信收发机等。通信单元809允许设备800通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0093] 计算单元801可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元801的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元801执行上文所描述的各个方法和处理,例如方法200-500。例如,在一些实施例中,方法200-500可被实现为计算机软件程序,其被有形地包含于机器可读介质,例如存储单元808。在一些实施例中,计算机程序的部分或者全部可以经由ROM 802和/或通信单元809而被载入和/或安装到设备800上。当计算机程序加载到RAM 803并由计算单元801执行时,可以执行上文描述的方法200-500的一个或多个步骤。备选地,在其他实施例中,计算单元801可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行方法200-500。

[0094] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0095] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处

理器或控制器，使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0096] 在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0097] 为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置(例如，CRT(阴极射线管)或者LCD(液晶显示器)监视器)；以及键盘和指向装置(例如，鼠标或者轨迹球)，用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈(例如，视觉反馈、听觉反馈、或者触觉反馈)；并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0098] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如，作为数据服务器)、或者包括中间件部件的计算系统(例如，应用服务器)、或者包括前端部件的计算系统(例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如，通信网络)来将系统的部件相互连接。通信网络的示例包括：局域网(LAN)、广域网(WAN)和互联网。

[0099] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

[0100] 应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本发公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本公开公开的技术方案所期望的结果，本文在此不进行限制。

[0101] 上述具体实施方式，并不构成对本公开保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等，均应包含在本公开保护范围之内。

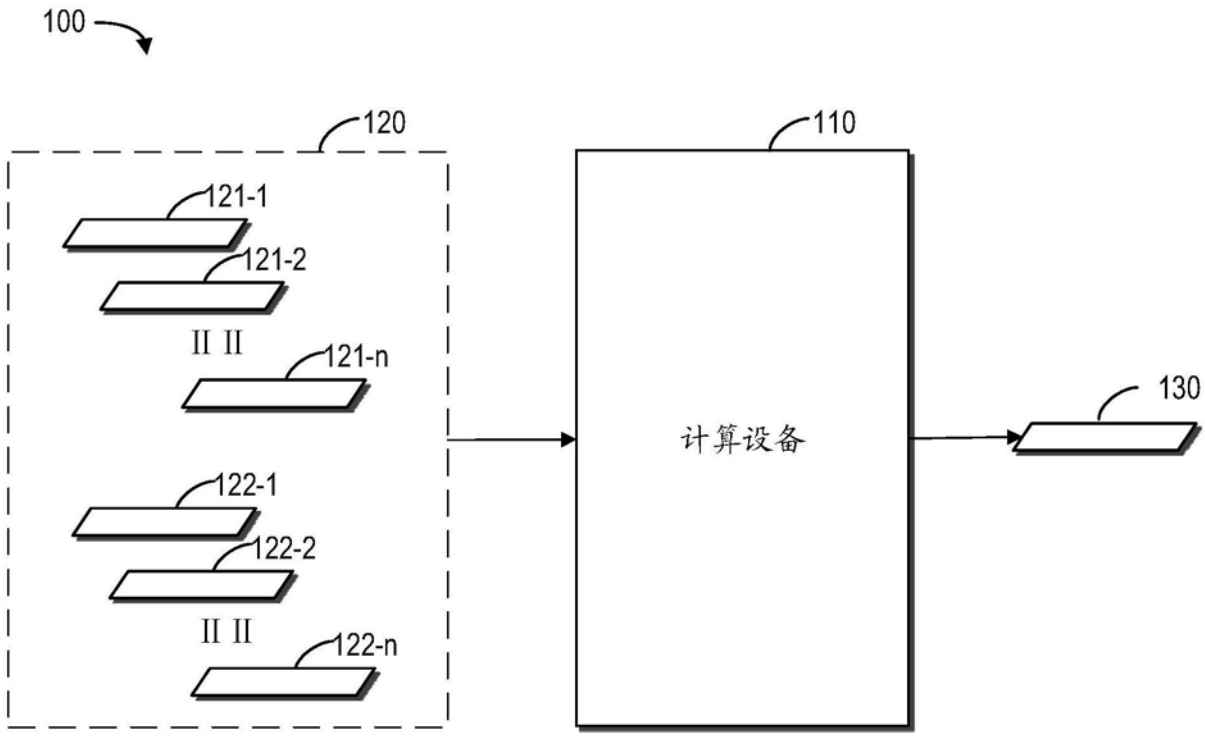


图1

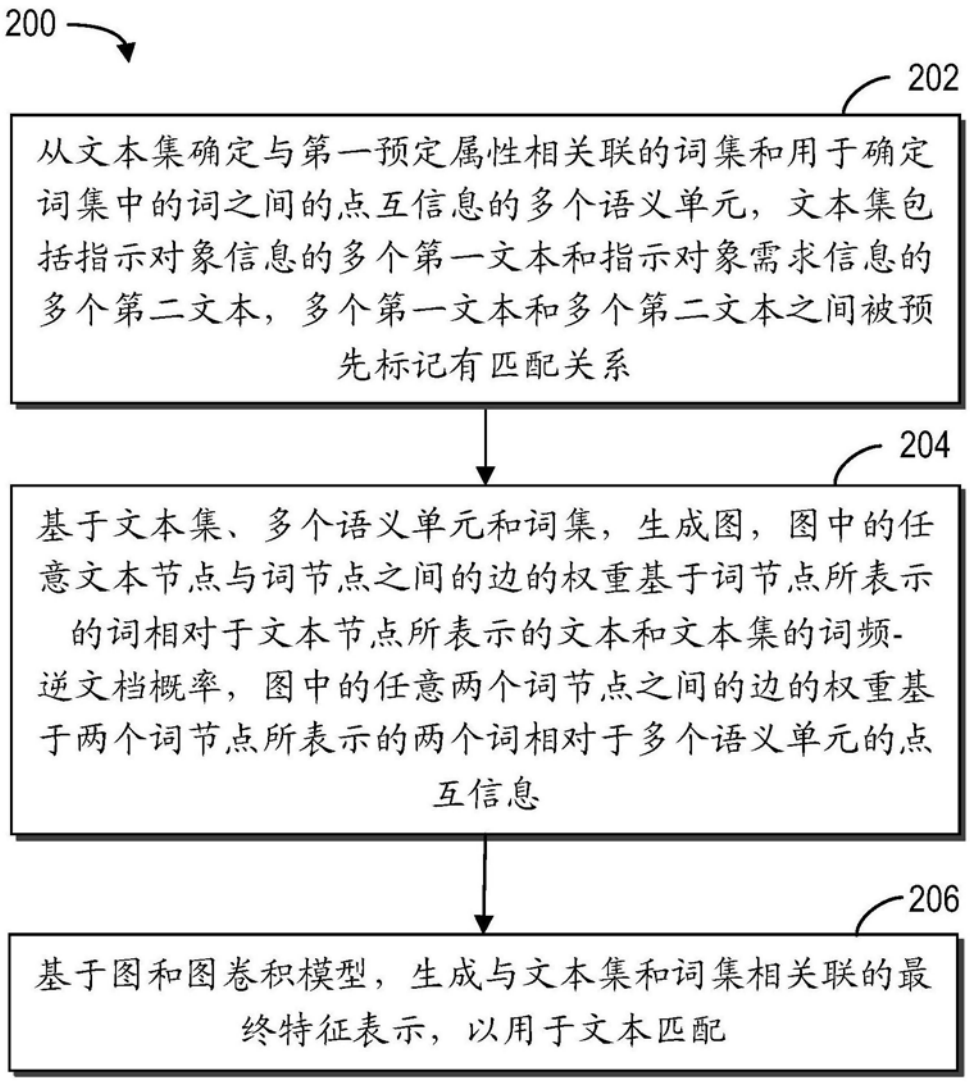

200
202
从文本集确定与第一预定属性相关联的词集和用于确定词集中的词之间的点互信息的多个语义单元，文本集包括指示对象信息的多个第一文本和指示对象需求信息的多个第二文本，多个第一文本和多个第二文本之间被预先标记有匹配关系
204
基于文本集、多个语义单元和词集，生成图，图中的任意文本节点与词节点之间的边的权重基于词节点所表示的词相对于文本节点所表示的文本和文本集的词频-逆文档概率，图中的任意两个词节点之间的边的权重基于两个词节点所表示的两个词相对于多个语义单元的点互信息
206
基于图和图卷积模型，生成与文本集和词集相关联的最终特征表示，以用于文本匹配

图2

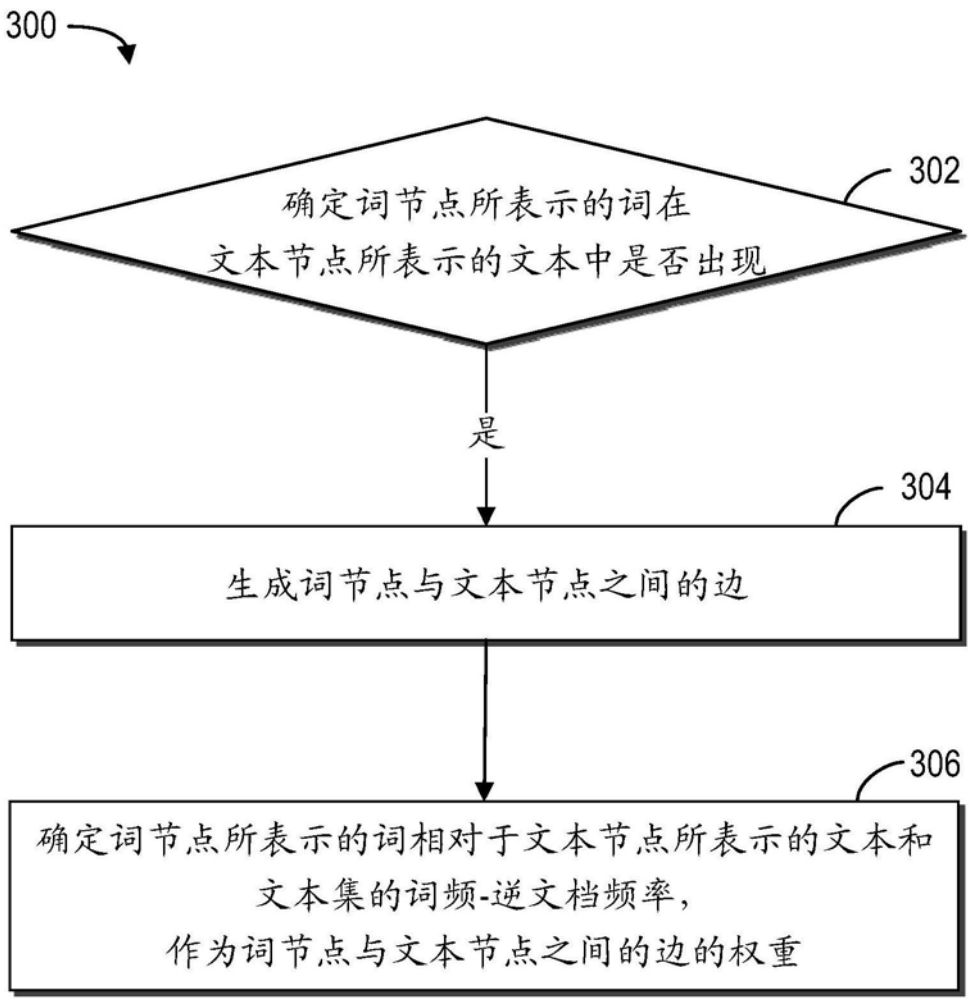

300
确定词节点所表示的词在
文本节点所表示的文本中是否出现
302
是
304
生成词节点与文本节点之间的边
306
确定词节点所表示的词相对于文本节点所表示的文本和
文本集的词频-逆文档频率，
作为词节点与文本节点之间的边的权重

图3

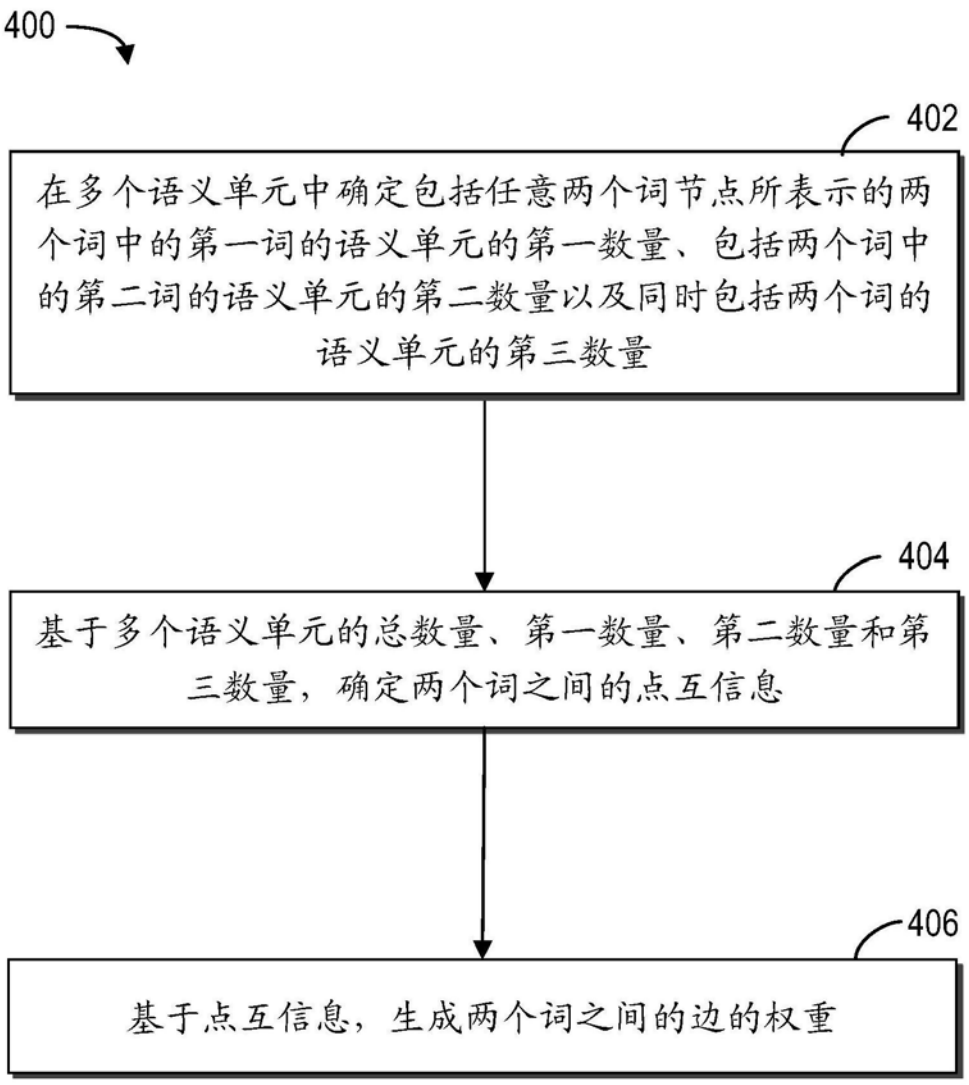

400
402
在多个语义单元中确定包括任意两个词节点所表示的两个词中的第一词的语义单元的第一数量、包括两个词中的第二词的语义单元的第二数量以及同时包括两个词的语义单元的第三数量
404
基于多个语义单元的总数量、第一数量、第二数量和第三数量，确定两个词之间的点互信息
406
基于点互信息，生成两个词之间的边的权重

图4

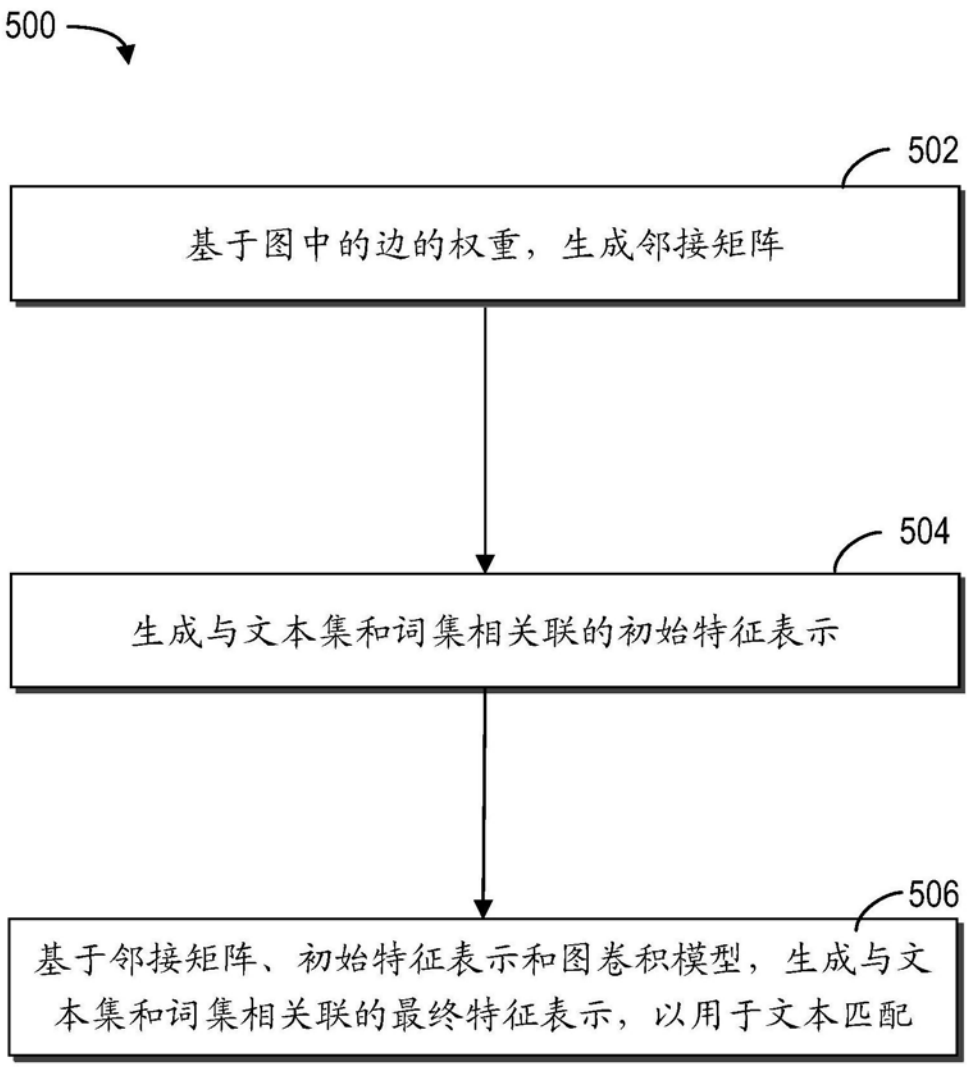

500
502
基于图中的边的权重，生成邻接矩阵
504
生成与文本集和词集相关联的初始特征表示
506
基于邻接矩阵、初始特征表示和图卷积模型，生成与文本集和词集相关联的最终特征表示，以用于文本匹配

图5

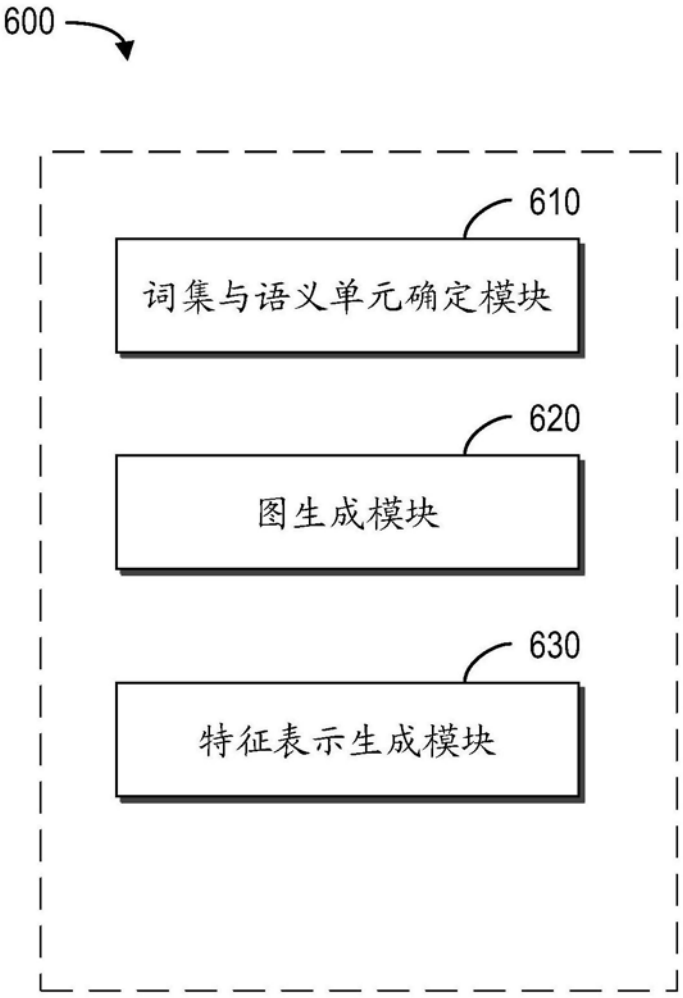


图6

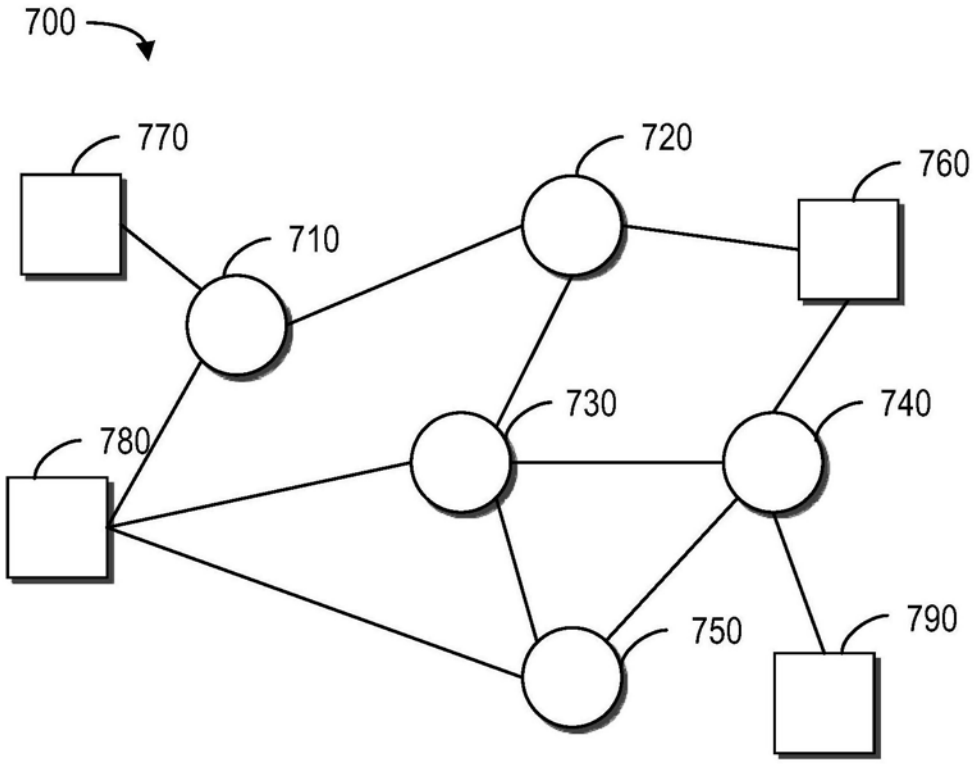

700
770
720
760
710
730
740
780
750
790

图7

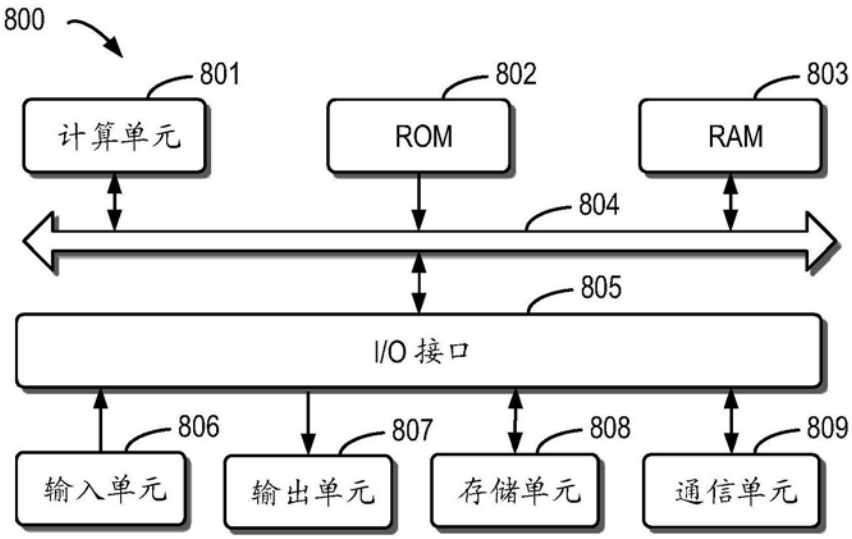

800
801
802
803
计算单元
ROM
RAM
804
805
I/O 接口
806
807
808
809
输入单元
输出单元
存储单元
通信单元

图8